『国語研日本語ウェブコーパス』からの新規語彙素獲得の試み

著者	岡 照晃
雑誌名	言語資源活用ワークショップ発表論文集
巻	3
ページ	586-592
発行年	2018
URL	http://doi.org/10.15084/00001693

『国語研日本語ウェブコーパス』からの新規語彙素獲得の試み

岡 照晃 (国立国語研究所コーパス開発センター) *

An Attempt to Extract New Lemma Candidates from Ninjal Web Japanese Corpus

Teruaki Oka (National Institute for Japanese Language and Linguistics)

要旨

『国語研日本語ウェブコーパス（NWJC）』は、国立国語研究所がこれまで公開してきた『現代日本語書き言葉均衡コーパス（BCCWJ）』や『日本語話し言葉コーパス（CSJ）』と異なり、形態論情報をすべて形態素解析器『MeCab』と『解析用 UniDic』を使って自動付与している。『BCCWJ』や『CSJ』といった既存のコーパスの整備の際には、コーパスアノテーションと同時に、形態論情報のデータベースである『UniDic DB』に新規短単位語彙素を追加していた。そのためコーパス整備と同時に『UniDic DB』も拡張されてきたが、『NWJC』は全自動で構築されたため、新規短単位語彙素の検出と DB への登録が行われておらず、その箇所で自動解析誤りのままとなっている。そこで本研究では、形態素解析を介さず、文字 N-gram の出現頻度と連接頻度の情報から文字 N-gram の分散表現を作成し、『NWJC』から『UniDic DB』に未登録の新規短単位語彙素の候補を列挙する方法について述べる。これにより DB のさらなる拡張が望めるだけでなく、『UniDic DB』のエクスポートデータで作成される『解析用 UniDic』も拡張されるため、それを用いた再解析によって『NWJC』中の誤解析箇所を減らすことにもつながる。

1. はじめに

国立国語研究所コーパス開発センターでは、現在、現代日本語の形態素解析用辞書として、現代書き言葉解析用 UniDic と現代話し言葉解析用 UniDic の 2 つを公開している [(1)]。これらの解析用辞書の語彙は共通であり、1,745,957 の書字形出現形（表層系）と、その活用変化や異語形を束ねた語彙素（辞書の見出し語相当）を 258,550 を含んでいる (表 1)。しかしながら、2013 年以来、（現代語の）解析用 UniDic には新規の短単位登録がなく、辞書の語彙もすでに古くなっている。これは解析用 UniDic の元となる短単位格納 DB、UniDic DB がコーパスアノテーションと同時に拡張されていくものであり、現代語のコーパス整備が現代日本語書き言葉均衡コーパス以来、大規模に行われなかったことによる。そこでコーパス開発センターでは現在、UniDic へ新たに 5,000 の新規語彙素の追加を計画している。その一環として、本研究では、

*teruaki-oka {at} ninjal.ac.jp

(1) http://unidic.ninjal.ac.jp/download/#unidic_bccwj

表 1　解析用 UniDic の語彙の統計。

語彙素数	258,550
書字形出現形数	1,745,957

表 2　NWJC の統計。

No. of URLs	83,992,556
Tokens	3,885,889,575
Types	1,463,142,939
No. of Characters	33,226,333,292

国語研日本ウェブコーパス（NWJC） [Asahara et al., 2014] からの新規語彙素の候補の抽出に取り組みについて述べる。日本語は英語など分かち書きする言語と異なり、単語境界がスペースで明示されない。そのため新規語彙素を文字列中から発見することは難しい。そこで本研究では、単語分割を介さずに文字 N-gram の分散表現ベクトルを学習する sembei [Oshikiri, 2017] というアルゴリズムを採用し、K-neighbour classification によって UniDic に既に登録されている短単位に近い分散表現ベクトルを持った文字 N-gram を新規語彙素の候補として抽出する。本手法により、名詞に関して精度よく新規語彙素の獲得が可能であることが分かった。

2. 国語研日本語ウェブコーパス (NWJC)

国語研日本語ウェブコーパスは、国語研コーパス開発センターで開発された 100 億語規模の日本語のウェブコーパスである。ウェブページの収集には Heritrix crawler [(2)]が使用され、このクローラを 1 億 URL について URL のリストを更新しつつ、3 か月に 1 度動かし、1 年間変化のないウェブページを収集した。クロールされたページは nwc-toolkit-0.0.2 [(3)]によって正規化（HTML タグ削除と NFKC 正規化の後、文へと分割）した。ウェブ上にはコピーされたページも存在し、収集したページの中にもそれは含まれている。そのため Unix の uniq コマンドを使用し、文を延べではなく、異なりにする作業を行い、重複を排除した。NWJC には、2014 年の 10〜12 月 (2014-4Q) に収集されたウェブページのデータが異なり文集合として格納されている。NWJC の統計データを表 2 に示す。

3. 関連研究

日本語の特徴の一つに分かち書きをしないことがあげられる。そのため日本語解析の一番最初のステップは、単語分割、品詞タグ付け、活用推定等を含んだ（日本語）形態素解析処理である。日本語形態素解析では、形態素解析用辞書を用いた手法が主流であり、広く使われているツールとして、CRF [Lafferty et al., 2001] で辞書に登録されている各表層形のコストを学習

[(2)] http://webarchive.jira.com/wiki/display/Heritrix/Heritrix/
[(3)] http://code.google.com/p/nwc-toolkit/

する MeCab [Kudo et al., 2004] [4]がある。

分かち書きされていない生文から新規語彙素の候補を列挙したいと思った場合、もっとも単純な方法は MeCab のような辞書ベースの形態素解析器を利用することである。MeCab をはじめ、辞書ベースの形態素解析器には未知語処理の機能が実装されており、解析結果が「未知語」となった文字列を切り出せば、それを新規語彙素の候補とできる。しかしこの方法の欠点はそもそも未知語が辞書に載っていないため、正しく切り出されないという点にある。また特に今回対象とする Web テキストにはくだけた表現や省略形も多く、それが辞書中の別の短いエントリとマッチして、未知語が未知語と認識されないことも多い。

そこで本研究では、形態素解析器を介さず、文字 N-gram の分散表現ベクトルを学習する sembei アルゴリズム [Oshikiri, 2017] を採用し、新規語彙素の候補を品詞ごとに NWJC から K-neighbour classifier を使って抽出する方法をとる。sembei アルゴリズムについては次節で述べる。本手法は単語の分散表現ベクトルと近傍法を利用したものであるが、本質的には [Mori et al., 1996] の文字ベースの日本語未知語獲得手法の亜種であるといえる。

4. 提案手法： sembei アルゴリズムと K-neighbour classifier を用いた新規語彙素候補の抽出

4.1 sembei アルゴリズムの詳細とパラメータ設定

sembei アルゴリズムのフレームワークでは、まずコーパス中に頻出する文字 N-gram で文をラティスに変換する。これは頻出文字 N-gram に基づく当該文の可能な分割候補を列挙しているとも捉えられるため、辞書ベースの形態素解析で構築される単語ラティスに類似のものである。次に N-gram ラティス上での共起（連接）の統計値を使い、N-gram のベクトルを学習する。

大規模なウェブコーパスである NWJC から頻出する N-gram を獲得するため、sembei では lossy counting アルゴリズムを使って低頻度要素の逐次削除と、頻度の近似値計算を採用している。本手法では、N-gram 長 : N= 1 〜8 を採用し、lossy counting アルゴリズムにより、NWJC から 22,455,810 個、長さ 1〜8 文字の異なり N-gram を獲得した。

sembei アルゴリズムでは、コーパス中の頻出文字 N-gram のみでラティスを構築する。これに対し本手法では、生コーパス中の高頻度 N-gram だけでなく、なるべく均等に N-gram を選択したい。そのため lossy counting アルゴリズムで獲得した全 N-gram の（近似）頻度の分布に基づき、ラティス構築のための N-gram をランダムに選択した。実際には、22,455,810 個の N-gram から 1,150,000 個を頻度の分布に基づいてランダムに抽出し、NWJC 中の文を N-gram ラティスに変換した。

次に、変換した N-gram ラティス上での N-gram 同士の共起（連接）頻度から各 N-gram のベクトルを学習する。Oshikiri (2017) では、negative sampling ありの skip-gram モデル (SGNS-sembei) [5]を採用しているが、ここではもともとの sembei アルゴリズム [6]が採用している

[4] https://taku910.github.io/mecab/
[5] https://github.com/oshikiri/w2v-sembei
[6] https://github.com/shimo-lab/sembei

eigenwords (OSCCA) [Dhillon et al., 2015] を使用した。ここで 1,150,000 個の N-gram 集合を $V = \{v_1, v_2, ..., v_{1150000}\}$ と表す。v_i は各 N-gram を表している。v_i の頻度を $\#(v_i)$ と表記し、v_i と v_j の連接頻度を $\#(v_i, v_j)$ と表記する。C_L は v_i に対する左連接行列で、C_R は v_i に対する右連接行列である。連接頻度のカウントにも、再度 lossy counting アルゴリズムを使用する。

$$C_L := \left(\frac{\#(v_j, v_i)}{\sqrt{\#(v_i)}\sqrt{\Sigma_k \#(v_j, v_k)}} \right)_{i,j} \in \mathbb{R}^{V \times V}$$

$$C_R := \left(\frac{\#(v_i, v_j)}{\sqrt{\#(v_i)}\sqrt{\Sigma_k \#(v_k, v_j)}} \right)_{i,j} \in \mathbb{R}^{V \times V}$$

$$C := [C_L, C_R]$$

ここで OSCCA の分散表現は $G_1^{-1/2}[\mathbf{u}_1, ..., \mathbf{u}_K]$ であり、G_1 は $\#(v_1), \#(v_2), ..., \#(v_V)$ を要素とする V 次対角行列である。$\mathbf{u}_1, ..., \mathbf{u}_K$ は $\sqrt{C}$ の左特異ベクトルの上位 K 件である。$\sqrt{C}$ の要素は $\sqrt{c}$ である。c は C の要素である。$\mathbf{u}_1, ..., \mathbf{u}_K$ を計算するため、randomized SVD [Halko et al., 2011] を使用し、$K = 200$ に設定した。この方法で、1,150,000 個の N-gram から、それぞれ 200 次元の 1,150,000 個のベクトル（分散表現）が作成された。

4.2 N-gram の分散表現からの新規語彙素候補の抽出

N-gram の分散表現から新規語彙素の候補を抽出するため、 Python machine learning toolkit である scikit-learn の K-Neighbour Classifier を使用する。K-Neighbour Classifier のコンストラクタ引数は、*n_neighbours* = 5、*weights* = '*distance*' に設定した。

N-gram($\in V$) がすでに書字形出現形として UniDic に登録されており、それが指定された品詞をただ 1 つを取りうる場合に限り、当該 N-gram の分散表現ベクトルを訓練用の正例とし、それ以外を訓練用の負例とする。また N-gram ($\in V$) が句読点のような記号を中に含む場合も、訓練用の負例とする。訓練用の N-gram を除いた V（UniDic に未登録の表層形）に対し、それが正例（新規語彙素の候補）か負例かを訓練用データで学習した K-neighbour classifier を使い判別する。

5. 結果：抽出された新規語彙素の候補

訓練用データの正例ので指定する品詞に名詞を設定したとき、抽出された結果はほとんどが名詞か名詞句であった。ただし、この抽出結果をそのまま UniDicDB に追加することはできない。UniDic の短単位登録には厳格な規則があり、DB への登録には人手の確認作業が必要になる。人手での確認作業を含めた提案手法を一度実施した結果、約 50 個の新規語彙素候補の獲得に成功した。品詞：名詞と指定して抽出した結果のうち、K-neighbour classifier の確信度（(K-neighbour classifier.pred_proba(X) 関数の出力)0〜1 の値をとる）で 1.0 で正例と判断した結果を表 3 に示す。

これに対し、品詞：動詞を指定した場合、抽出された候補 N-gram のほとんどが副詞 (e.g. たっぷり, たくさん）＋動詞（UniDic DB に登録済み）であった。sembei アルゴリズムは左右に連接する N-gram の情報のみから分散表現ベクトルを作成するため、当該の N-gram である

表 3　正例と予測され抽出された新規語彙素候補の例（名詞）

新規語彙素候補	説明
ＵＦＣ	Ultimate Fighting Championship
Ｗｉ－Ｆｉ	
ＷｉＭＡＸ	
ＹｏｕＴｕｂｅ	
ｉＭａｃ	
ｉＰａｄ	
ｉＰｈｏｎｅ	
ｉＰｏｄ	
ｉＴｕｎｅｓ	
ｐｉｘｉｖ	
ｔｗｉｔｔｅｒ	
のだめ	キャラクター名
ネトゲ	
ホスホニウム	化学用語（Phosphonium）
ホルミル	化学用語（Formyl）
マツコ	人名
ラジコ	サービス名
ルキア	キャラクター名
絵茶	お絵描きチャット
電マ	電動マッサージ
播戸	苗字

v_i の情報は、分散表現ベクトルの学習時に一切使用しないことが原因だと考えられる。そこで例えば、UniDic に既に登録済みの部分文字列を含む N-gram をあらかじめ除外する、もしくは部分文字列として含むという事柄を学習時に素性として追加することが考えられる。

品詞：形容詞を指定したとき、事前の調査で発見していた NWJC 中の新規語彙素「エモい」は獲得されなかった。これは NWJC 中に「エモい」の出現がわずか 500 件しかなく、lossy counting の時点で頻出 N-gram から漏れていたことが原因である。lossy counting のパラメータを調整したところ、「エモい」も頻出 N-gram として残ったが、頻出 N-gram 数が 7 億と非常に巨大なサイズになった。そのため今後は、OSSCA による行列のバッチ計算から SGDN を使ったオンライン学習による分散表現学習に切り替える方針である。

6. おわりに

本稿では、NWJC のような巨大なウェブコーパスから UniDic の新規語彙素候補をどのようすれば獲得できるのかを示した。sembei アルゴリズムを採用し、文字 N-gram の分散表現を学

表 4　正例と予測され抽出された新規語彙素候補の例（動詞）

新規語彙素候補
たくさんし
たくさんつい
たくさん飲
たくさん作っ
たくさん出
たくさん書い
たくさん入れ
たくさん落ち
たっぷり含まれ
たっぷり詰まっ
たっぷり使
たっぷり使っ
たっぷり入っ

習したのち、K-neighbour classification によって文字 N-gram の中から新規語彙素候補を識別した。名詞に関しては新規語彙素の候補が獲得できたが、動詞に対しては副詞を含むといった冗長抽出が起きてしまった。これを避けるには、文字列前方に副詞を含むような N-gram をあらかじめ除外、もしくは負例にするなどの方法が検討される。

謝　辞

本研究は国立国語研究所コーパス開発センターの共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」(2016-2021 年度) の成果である。

文　献

[Asahara et al., 2014] Asahara, M., Maekawa, K., Imada, M., Kato, S., and Konishi, H. (2014). Archiving and Analysing Techniques of the Ultra-large-scale Web-based Corpus Project of NINJAL, Japan *Alexandria*, 25(1-2):129–148.

[Bojanowski et al., 2016] Bojanowski, P., Grave, E. Joulin, A and Mikolov, T. (2016). *arXiv* preprint arXiv:1607.04606.

[Dhillon et al., 2015] Dhillon, P. S., Foster, D. P. and Ungar, L. H. (2015). Eigenwords: Spectral word embeddings. *Journal of Machine Learning Research*, 16: 3035–3078.

[Kono et al., 2015] Kono, T. and Ogiso, T. (2015). Improving an Electronic Dictionary for Morphological Analysis of Japanese: Use of historical period information *Proceedings of ICML-2001 (the 18th International Conference on Machine Learning)*, pages 282–289.

[Mank., 2002] Manku, G. S. and Motwani, R. (2002). Approximate frequency counts over data

streams. *Proceedings of VLDB '02*, pages 346–357.

[Halko et al., 2011] Halko, N., Martinsson, P. G. and Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2): 217–288.

[Kudo et al., 2004] Kudo, T., Yamamoto, K. and Matsumoto, Y. (2004). Applying Conditional Random Fields to Japanese Morphological Analysis. *Proceedings of EMNLP-2004 (the 2004 Conference on Empirical Methods in Natural Language Processing)*, pages 230–237.

[Lafferty et al., 2001] Lafferty, J., McCallum, A. and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of ICML-2001 (the 18th International Conference on Machine Learning)*, pages 282–289.

[Mori et al., 1996] Mori, S. and Nagao, M. (1996). Word Extraction from Corpora and Its Part-of-Speech Estimation Using Distributional Analysis. *Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996)*, pages 1119–1122.

[Oshikiri, 2017] Oshikiri, T. (2017). Segmentation-Free Word Embedding for Unsegmented Languages. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 778–783.